Data Mining Lab 2020: Variational Fair Autoencoder

Mohammad Abir, Fabian Beringer

Background

The Fair Classification Setting

- feature vectors **x** contain protected attribute **s**
- **s**=1 indicates membership in the protected group, **s**=0 no membership
- for example: *gender, ethnicity, age, religion*
- just removing **s** is not sufficient because of *proxies* i.e. **s** could be inferred from the other attributes

The Fair Classification Setting

- One way to measure fairness:

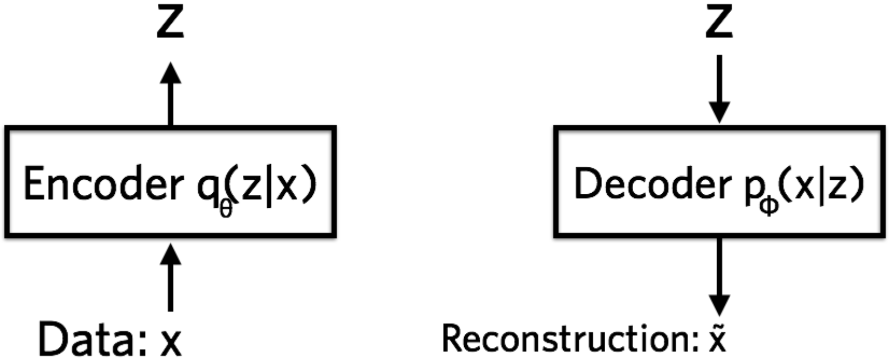
  Group fairness / statistical parity
  The *ratio of positive predictions* in the *protected group* (i.e. the instances with **s**=1) should be equal to the ratio for the non-protected group (instances where **s**=0)

Variational Autoencoders

- probabilistic framework for learning low dimensional representations
- usually implemented using neural networks
- an *encoder* and a *decoder* network try to generate and reconstruct data respectively

Variational Autoencoders

- encoder models a probability distribution over latent representations **z** given data inputs **x**
- decoder tries to reconstruct **x** from **z**



source: https://jaan.io/what-is-variational-autoencoder-vae-tutorial/

Variational Autoencoders

- to approximate $p(z \mid x)$, use KL-divergence
- **KL(** $q_\phi(z \mid x) \;\|\; p(z \mid x)$ **)** = $\mathbf{E}_q$**[** $\log q_\phi(z \mid x)$ **]** - $\mathbf{E}_q$**[** $\log p(x, z)$ **]** + $\log p(x)$

  problem**:** $p(x)$ is intractable!

- instead, optimize lower bound:

  $\mathbf{E}_q$**[** $\log p(x, z)$ **]** - $\mathbf{E}_q$**[** $\log q_\phi(z \mid x)$ **]**

Problem Statement

- the goal is to find useful representations for a set of data points
- data contains one or more random variables that are sensitive, i.e. they're prone to discrimination
- the learned representations should be:
  - invariant of the sensitive variables
  - contain as much information as possible for downstream tasks, e.g classification or clustering

Methodology

Unsupervised approach

- based on the *Variational Autoencoder* architecture
- It models a sensitive variable **s** and a latent variable **z** as independent sources of **x**

$$\mathbf{z} \sim p(\mathbf{z}); \qquad \mathbf{x} \sim p_\theta(\mathbf{x} \mid \mathbf{z}, \mathbf{s})$$

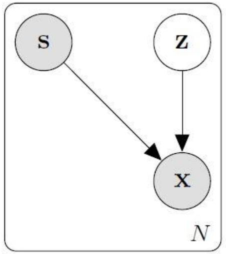


Figure 1: Unsupervised model

Semi-Supervised Approach

- However in cases where **s** and labels **y** are are correlated, **z** might become *random* or *degenerate* with respect to **y**
- Solution: Try to correlate **y** and **z** by injecting **y** into **z**
- This results in two latent variables $\mathbf{z}_1$ and $\mathbf{z}_2$

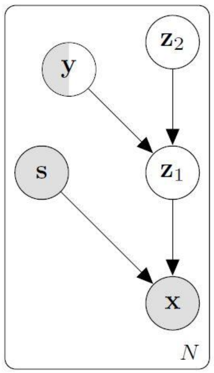


Figure 2: Semi-supervised model

Similar to the original VAE, a lower bound is optimized:

$$\sum_{n=1}^{N} \log p(\mathbf{x}_n|\mathbf{s}_n) \geq \sum_{n=1}^{N} \mathbb{E}_{q_\phi(\mathbf{z}_{1n},\mathbf{z}_{2n},\mathbf{y}_n|\mathbf{x}_n,\mathbf{s}_n)}[\log p(\mathbf{z}_2) + \log p(\mathbf{y}_n) + \log p_\theta(\mathbf{z}_{1n}|\mathbf{z}_{2n},\mathbf{y}_n) +$$
$$+ \log p_\theta(\mathbf{x}_n|\mathbf{z}_{1n},\mathbf{s}_n) - \log q_\phi(\mathbf{z}_{1n},\mathbf{z}_{2n},\mathbf{y}_n|\mathbf{x}_n,\mathbf{s}_n)] \quad (2)$$

Maximum Mean Discrepancy

- To further disentangle **z** and **s**, a penalty term based on *Maximum Mean Discrepancy* is added to the objective function
- This enforces a matching between the moments of the distributions over $\mathbf{z}_1$ for $\mathbf{s} = 0$ and $\mathbf{s} = 1$
- To shorten computation time they use the *Fast MMD* implementation of *MMD* which is an approximation

$$\ell_{\mathrm{MMD}}(\mathbf{Z}_{1\mathbf{s}=0}, \mathbf{Z}_{1\mathbf{s}=1}) = \| \mathbb{E}_{\tilde{p}(\mathbf{x}|\mathbf{s}=0)}[\mathbb{E}_{q(\mathbf{z}_1|\mathbf{x},\mathbf{s}=0)}[\psi(\mathbf{z}_1)]] - E_{\tilde{p}(\mathbf{x}|\mathbf{s}=1)}[\mathbb{E}_{q(\mathbf{z}_1|\mathbf{x},\mathbf{s}=1)}[\psi(\mathbf{z}_1)]]\|^2$$

$$\psi_{\mathbf{W}}(\mathbf{x}) = \sqrt{\frac{2}{D}} \cos\left(\sqrt{\frac{2}{\gamma}}\mathbf{x}\mathbf{W} + \mathbf{b}\right)$$

Evaluation

- The evaluation was performed by training a *random forest* and a *logistic regression* model on the VFAE obtained representations
- Accuracy for predicting **y** and **s** is measured, as well as two discrimination metrics:

$$\text{Discrimination} = \left| \frac{\sum_{n=1}^{N} \mathbb{I}[y_n^{s=0}]}{N_{s=0}} - \frac{\sum_{n=1}^{N} \mathbb{I}[y_n^{s=1}]}{N_{s=1}} \right|$$

$$\text{Discrimination prob.} = \left| \frac{\sum_{n=1}^{N} p(y_n^{s=0})}{N_{s=0}} - \frac{\sum_{n=1}^{N} p(y_n^{s=1})}{N_{s=1}} \right|$$

DESCRIPTION OF DATASETS

	Prediction Criteria/ Dataset Info	Total No of Data Points	Sensitive Variable (**s**)	Source
GERMAN	If Person has Good/Bad Credit	1000	Gender of Individual FEMALE - 310 (ProtectedClass) MALE - 690 (UnprotectedClass)	UCI machine learning repository (Frank & Asuncion, 2010)
ADULT INCOME	Predict Account holder has over 50,000 dollars in their account	**45,222**	Age Age > 65 = **1803** (ProtectedClass) Age < 65 = **47039** (UnprotectedClass)	UCI machine learning repository (Frank & Asuncion, 2010)
HEALTH	Predict whether a patient will spend any days in the hospital in the next year	147, 473	Age of Individual	Heritage Health Prize www.heritagehealthprize.com

DESCRIPTION OF DATASETS

EXTENDED YALE B	Face images of 38 people under different lighting conditions (directions of the light source)		5 states: **Light source** in: upper right, lower right, lower left, Upper left, Front.	Employed by Li et al. (2014).
AMAZON REVIEWS	Text Reviews about Products belonging to domains: ““books”, “dvd”, “electronics” and “kitchen”			Employed by Chen et al. (2012) and Ganin et al. (2015)

DESCRIPTION OF DATASETS

Fairness Representation | **Domain-Adaptation** | **Invariant Representation**

Fairness Representation

-- German, Adult Income and Health Dataset were used
-- Data was Binarized
-- A multivariate Bernoulli distribution was used where σ(·) is the sigmoid function.

$$p_\theta(\mathrm{x}_n|\mathrm{z}_{1n}, \mathrm{s}_n) = \mathrm{Bern}(\mathrm{x}_n|\pi_n = \sigma(f_\theta(\mathrm{z}_{1n}, \mathrm{s}_n)))$$

Domain-Adaptation

-- Amazon Reviews Dataset were used.
-- 12 Tasks were completed
-- Label '**y**' is correspended to each sentiment (Pos or Neg)
-- Poisson Distribution is used as each feature vector **x** is composed from counts of unigrams and bigrams

$$p_\theta(\mathbf{x}_n|\mathbf{z}_{1n}, \mathbf{s}_n) = \mathrm{Poisson}(\mathrm{x}_n|\lambda_n = e^{f_\theta(\mathbf{z}_{1n}, \mathbf{s}_n)})$$

Invariant Representation

-- Extended Yale B Dataset were used.
-- 12 Tasks were completed
-- Label '**y**' is corresponded to each sentiment (Pos or Neg)
-- Poisson Distribution is used
-- For the distribution, a Gaussian with means constrained in the 0-1 range (since we have intensity images) by a sigmoid

$$p_\theta(\mathbf{x}_n|\mathbf{z}_{1n}, \mathbf{s}_n) = \mathcal{N}(\mathrm{x}_n|\mu_n = \sigma(f_\theta(\mathrm{z}_{1n}, \mathrm{s}_n)), \sigma_n = e^{f_\theta(\mathbf{z}_{1n}, \mathbf{s}_n)})$$

Experimental Setup

- Latent dimensions - 50 except German Dataset - 30
- Simple Logistic Regression Classifier is used
- Optimization hyperparameters:
  - Adam with default settings, minibatch size 100

Experimental Setup

- Scaling Lower Penalty done by - MMD Penalty X Minibatch Size of 100
- MMD, **β** Tuned according to Validation Set
- Scaling of supervised cost :

  $\alpha = 1$ for the Adult, Health and German dataset,

  $\alpha = 100$ for Amazon Reviews(empirically determined),

  $\alpha = 200$ for Yale B Dataset
- Scaling of the MMD penalty was :

  For the Amazon reviews dataset $\beta = 100 \cdot N_{batch}$

  For the Extended Yale B $\beta = 200 \cdot N_{batch}$

Experimental Setup

- Classification Performance on **y** :
  VAE/VFAE: Predictive Posterior $q_\phi(\mathbf{y}|\mathbf{z}_1)$
  The original Representations **x**: Simple Logistic Regression
- K = 50 for Latent Space as Baseline for Learning Fair Representation
- Accuracy Measurement in '**y**' by LFR Model Predictions
- The discrimination metric used from Zemel et al. (2013) and updated version of the discrimination metric (Taking account of Probabilities of the correct class)